

**Listing of the Claims:**

1. (original) A method for clustering a string, the string including a plurality of characters, the method including:

identifying R unique n-grams  $T_{1\dots R}$  in the string;

for every unique n-gram  $T_S$ :

if the frequency of  $T_S$  in a set of n-gram statistics is not greater than a first threshold:

associating the string with a cluster associated with  $T_S$ ;

otherwise:

for every other n-gram  $T_V$  in the string  $T_{1\dots R}$ , except S:

if the frequency of n-gram  $T_V$  is greater than the first threshold:

if the frequency of n-gram pair  $T_S-T_V$  is not greater than a second threshold:

associating the string with a cluster associated with the n-gram pair  $T_S-T_V$ ;

otherwise:

for every other n-gram  $T_X$  in the string  $T_{1\dots R}$ , except S and V:

associating the string with a cluster associated with the n-gram triple  $T_S-T_V-T_X$ ;

otherwise:

do nothing.

2. (original) The method of claim 1 further including compiling n-gram statistics.

3. (original) The method of claim 1 further including compiling n-gram pair statistics.
4. (original) A method for clustering a plurality of strings, each string including a plurality of characters, the method including:
  - identifying unique n-grams in each string;
  - associating each string with clusters associated with low frequency n-grams from that string, if any; and
  - associating each string with clusters associated with low-frequency pairs of high frequency n-grams from that string, if any.
5. (original) The method of claim 4 further including:
  - where a string does not include any low-frequency pairs of high frequency n-grams, associating that string with clusters associated with triples of n-grams including the pair.

6. (original) A method for clustering a string, the string including a plurality of characters, the method including:

identifying R unique n-grams  $T_{1\dots R}$  in the string;

for every unique n-gram  $T_S$ :

if the frequency of  $T_S$  in a set of n-gram statistics is not greater than a first threshold:

associating the string with a cluster associated with  $T_S$ ;

otherwise:

for  $i = 1$  to  $Y$ :

for every unique set of  $i$  n-grams  $T_U$  in the string  $T_{1\dots R}$ , except  $S$ :

if the frequency of the n-gram set  $T_S-T_U$  is not greater than a second threshold:

associating the string with a cluster associated with the n-gram set  $T_S-T_U$ ;

if the string has not been associated with a cluster with this value of  $T_S$ :

for every unique set of  $Y+1$  n-grams  $T_{UY}$  in the string  $T_{1\dots R}$ , except  $S$ :

associating the string with a cluster associated with the  $Y+2$  n-gram group  $T_S-T_{UY}$ .

7. (original) The method of claim 6 where  $Y = 1$ .

8. (original) The method of claim 6 further including compiling n-gram statistics.

9. (original) The method of claim 6 further including compiling n-gram group statistics.

10. (original) A computer program, stored on a tangible storage medium, for use in clustering a string, the program including executable instructions that cause a computer to:

identify R unique n-grams  $T_{1\dots R}$  in the string;

for every unique n-gram  $T_S$ :

if the frequency of  $T_S$  in a set of n-gram statistics is not greater than a first threshold:

associate the string with a cluster associated with  $T_S$ ;

otherwise:

for every other n-gram  $T_V$  in the string  $T_{1\dots R}$ , except S:

if the frequency of n-gram  $T_V$  is greater than the first threshold:

if the frequency of n-gram pair  $T_S-T_V$  is not greater than a second threshold:

associate the string with a cluster associated with the n-gram pair

$T_S-T_V$ ;

otherwise

for every other n-gram  $T_X$  in the string  $T_{1\dots R}$ , except S and V:

associate the string with a cluster associated with the n-gram triple  $T_S-T_V-T_X$ ;

otherwise:

do nothing.

11. (original) The computer program of claim 10 further including executable instructions that cause a computer to compile n-gram statistics.

12. (original) The computer program of claim 10 further including executable instructions that cause a computer to compile n-gram pair statistics.